Foreword

Intel® Xeon Phi™ Processor High Performance Programming: Knights Landing Edition

In their first *Intel® Xeon Phi™ Coprocessor High-Performance Programming* book, Jeffers and Reinders utilize a sports car analogy to introduce high performance computing. I like this analogy because I see many parallels between the automotive and computing worlds. There are drivers who see vehicles solely as appliances that get them from point A to point B. Perhaps these are people who look forward to self-driving cars. Similarly, there are computer users that are not concerned with performance–they are willing to wait however long is needed for their applications to complete. But if you are reading this book, you are likely very interested in computer performance. For you, a sports car analogy is appropriate for how you drive your computers. It is good to see that the authors' sports car introductory tutorial is available to an even wider audience at their open website, http://www.lotsofcores.com/sportscar.

Extending the sports car analogy to higher performance

I work at Sandia National Laboratories, and as with many of my colleagues at the U.S. Department of Energy national laboratories, we are interested in the highest possible performance. To carry forward the automotive analogy, we could say that as extreme performance supercomputer users, we are drivers of race cars. We not only want to reach our modeling and simulation goals as fast as possible but we want to look under the hood, understand what's there, and modify the race car to go even faster. It is also accurate to note that there is a larger population of national lab users that would be considered sports car drivers–they don't develop or modify the applications they use. They are interested in performance, but don't have the time to tune and optimize their applications for the highest possible performance. Instead, these HPC users are focused on their scientific research or engineering analysis so for them, supercomputers with their relevant modeling and simulation applications are tools to support their R&D.

I must also confess that I am an automotive enthusiast. I recall the thrill of receiving my driver's license, prefer driving a stick shift, and enjoy taking long road trips. I won't be a customer for the self-driving car. As an undergraduate mechanical engineer at the University of Illinois, I took every automotive engineering course that was offered. My vehicle dynamics class was taught by Professor Robert A. White, who also consulted for the Porsche Research Center. This explains why I read with great interest, *The Unfair Advantage*[1], a book by Mark Donohue about his career driving and developing race cars. I especially enjoyed chapter 25, which describes the collaboration that Donohue and the Penske racing team established with Porsche on the development of the Type 917 race car.

What exactly is The Unfair Advantage?

Mark Donohue was not just a race car driver. He was also an automotive engineer who was able to translate what he felt through the throttle, brake pedal, steering wheel and racing seat into improvements to his race car. These could be simply the tuning adjustments that mechanics would make. Or these could be more drastic changes that were needed in the design of sub-systems that an engineer would make, e.g. changes to suspension geometry, or chassis bodywork and airflow. I was impressed by Donohue's following passage.

[1] Mark Donohue with Paul Van Valkenburgh, *The Unfair Advantage*, Robert Bentley, Inc., Cambridge, MA, USA, second edition, 2000.

> *We knew a lot about the engineering they were doing, and we had already come to some of the same conclusions. He [Helmut Flegl, the Porsche 917 chief engineer] had been engineering race cars for some years, and as he began to realize that we could relate, I could almost see a spark come to his eyes. There are certain things that are of interest only to racing engineers–like lateral acceleration in "g's," aerodynamic downforce, centers of mass–and we spoke the common language. We began to convince him that we were not like any other race team he had ever worked with.*[2]

The Penske racing team entered into an agreement with Porsche to collaborate directly on the development of the turbocharged version of the 917 race car when it was still a pre-production prototype. As a mechanical engineer, Donohue was able to communicate with the 917 race car designers in engineering terms. As a driver, Donohue had a first-hand understanding of the user requirements to refine and complete the final development of the 917. This was Donohue's *Unfair Advantage*.

Peak Performance versus Drivable/Usable Performance

In 1971 the Porsche engineers originally designed their 917 race car engine to produce maximum horsepower. Donohue and the Penske racing team were the first people outside of Porsche to receive this turbocharged engine. Unfortunately, according to Donohue, the 917 simply would not idle or run at part throttle. No amount of tuning by their mechanics could make the engine work. In the end, the Penske team had to stop testing the turbocharged engine in the 917 race car and they went back to the Porsche engine test stands and dynos. I pick up Donohue's description.

> *I looked at their dyno output curves. They had all the necessary data–torque, rpm, boost pressure, and so on–except that the curves started at 5,000 rpm. I said, "Why are there no curves up to that point?" They said, "The motor does not run there." I thought "Christ! That's what I've been trying to tell you for a month!" I couldn't believe it was that simple. I couldn't believe that they had simply calibrated the fuel injection for wide open throttle with full boost, and totally ignored any part-throttle operation. Flegl and I sat down and designed a dyno program to get the information we needed for proper calibration.*[3]

In short order, the turbocharged engine was properly calibrated to operate at all engine speeds and turbo boost pressure levels. This is the driver's perspective on *The Unfair Advantage*. Engineers can be misled by simple benchmarks that do not reflect how high performance systems are actually used. As a driver, Donohue needed the engine to operate well at all engine speeds from idle on up to redline. The Porsche engineers were not to blame for this oversight, they delivered what was asked for–an engine designed for maximum horsepower. They did not understand the user perspective because they were not drivers.

Let me segue back to the analogy between supercomputers and race cars. In the supercomputing community our baseline metric is High Performance Linpack (HPL) and the units of measure are floating point operations per second. This benchmark is used by the Top500 supercomputing sites list [4], and has been useful for providing a simple measure that can be used to characterize system performance. However, HPC users also understand that their real applications are usually not

[2] Ibid., page 282.
[3] Ibid., page 289.
[4] see http://top500.org

represented by how the HPL benchmark tests supercomputer capabilities. In recent years, a singular focus on HPL has led computer and system engineers to design supercomputers that can generate peak HPL benchmark measurements, but have not been very usable for real-world HPC applications. The concepts of useable and drivable are synonymous for high-performance systems whether they are supercomputers or race cars.

In the end, Donohue and the Penske racing team helped Porsche "complete" the engineering and development of their 917 race car, with key collaborative improvements in a variety of sub-systems from suspension geometry and engine tuning to vehicle aerodynamics. The bottom line is that as a driver, i.e. user of the race car system, Donohue had insights regarding how best to design and tune the system for optimal performance. But as an engineer, he was able to communicate how to improve the Porsche 917 design.

How does The Unfair Advantage relate to this book?

Section I is for readers that consider themselves to be race car "engineers/drivers." They are interested in extracting the highest performance potential of the underlying hardware. Within the national labs and in some university and commercial settings, these users are likely to be developers of architecture-centric software capabilities. For example, they may need to develop highly tuned and optimized math libraries that extract performance from all the architectural capabilities that were designed into Intel® Xeon Phi™ Knights Landing by Avinash Sodani and his processor architecture team. The audience for Section I may include users that are interested in the development of new system software to support the mapping of many HPC applications to Xeon Phi processor architectures. Finally, the audience for Section I may also be interested in how co-design can influence the design of future hardware.

Section II is for the race car and enthusiast sports car driver of Xeon Phi supercomputers. The chapters in this section describe how to program the Intel® Xeon Phi™ for those users that are interested in realizing the performance potential of the new Knight Landing architectural capabilities. These chapters are for readers that need to understand how to develop HPC applications to leverage the new architectural capabilities that are provided by the Xeon Phi Knights Landing. This audience may also be interested in how co-design principles are used to understand the tuning and application modifications needed to exploit the various advanced architecture features provided in the Xeon Phi Knights Landing processor.

Section III is for the sports car driver. This section of the book provides application examples with Intel® Xeon Phi™ Knights Landing results for quickly coming up to speed for a diverse portfolio of HPC applications. It may be possible to directly apply the lessons and patterns in these application examples to the reader's own applications, or the reader may already be a user of these applications in their own HPC workloads. Section III is also for readers that are interested in seeing examples for how the parallel programming concepts of Section II are implemented in real applications.

Closing Comments

Sandia National Laboratories is the DOE/National Nuclear Security Administration's (NNSA) engineering lab. I have often discussed with my colleagues that as an engineering lab, we are in a position to foster opportunities to collaborate with industry to help develop our needed supercomputers and supporting computing technologies. Recently, I helped write the NNSA Advanced Simulation and

Computing (ASC) program's Co-design Strategy.[5] This document describes different levels of co-design including *transformative* co-design as a way to influence future hardware designs.

While not every supercomputer user wants to help develop supercomputer technology, we have many scientists and engineers at Sandia's Center for Computing Research that are not afraid of driving first-of-a-kind, pre-production prototype computers. Yes, there is a risk of crashing, but the opportunity to drive leading edge computer systems is often one of the main attractions we can offer to our technical staff. The point is not just to have early access to hardware. These early testbeds foster direct collaborations between our adventurous "engineer-drivers" and our computer engineering/computer science counterparts in industry. These collaborative discussions are the payoff for being the first to drive new HPC technology. This is the "spark in the eye" described in Donohue's early conversation with Flegl when they realized they had a common engineering understanding.

I am grateful that my team at the Center for Computing Research has had similar conversations with Jim Jeffers and his Intel colleagues through our early access to three generations of pre-production versions of Intel® Xeon Phi™ processors. I would also like to acknowledge the long term collaboration that Sandia has established with support from many Intel executives including Raj Hazra, Charlie Wuischpard, Joe Curley, Thor Sewell, Mike Julier, Ranna Prajapati, Thomas Metzger, and Rajesh Agny to establish our NNSA/ASC-funded series of first-of-a-kind, Xeon Phi™ Knights Ferry, Knights Corner and Knights Landing rack-scale testbeds. My Sandia team includes James Laros III, Simon Hammond, Sue Kelly, Jim Brandt, and Ann Gentile; with strong management and programmatic support from Bruce Hendrickson, Ken Alvin, Rob Hoekstra, Tom Klitsner, and John Noe. These collaborations with Intel have helped us jointly develop lessons learned on our pre-production testbeds that we can go forward to apply on our DOE production Xeon Phi and Xeon supercomputers.

I believe you find this book is an invaluable reference to help develop your own *Unfair Advantage*!

James A. Ang, Ph.D.
Manager, Exascale Computing Program
Center for Computing Research
Sandia National Laboratories
Albuquerque, New Mexico USA
February 2016

[5] *ASC Co-design Strategy*, NNSA's Advanced Simulation and Computing Program, NA-114, February 2016, http://nnsa.energy.gov/sites/default/files/nnsa/inlinefiles/ASC_Co-design.pdf